

This is an electronic preprint version of the article

HEINZE, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data., *Statistics in Medicine* 25, 4216-4226.

Statistics in Medicine © 2006 John Wiley & Sons, Ltd.

<http://www.interscience.wiley.com/>.

A comparative investigation of methods for logistic regression
with separated or nearly separated data

Georg Heinze

Section of Clinical Biometrics, Core Unit for Medical Statistics and Informatics,

Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria

email: georg.heinze@meduniwien.ac.at

phone: +43-1-40400-6684

ABSTRACT

In logistic regression analysis of small or sparse data sets, results obtained by classical maximum likelihood methods cannot be generally trusted. In such analyses it may even happen that the likelihood meets the convergence criteria while at least one parameter estimate diverges to $\pm\infty$. This situation has been termed ‘separation’, and it typically occurs whenever no events are observed in one of the two groups defined by a dichotomous covariate. More generally, separation is caused by a linear combination of continuous or dichotomous covariates that perfectly separates events from non-events. Separation implies infinite or zero maximum likelihood estimates of odds ratios, which are usually considered unrealistic. I provide some examples of separation and near-separation in clinical data sets and discuss some options to analyze such data, including exact logistic regression analysis and a penalized likelihood approach. Both methods supply finite point estimates in case of separation. Profile penalized likelihood confidence intervals for parameters show excellent behavior in terms of coverage probability and provide higher power than exact confidence intervals. General advantages of the penalized likelihood approach are discussed.

KEY WORDS: bias reduction, exact logistic regression, infinite estimates, modified score function, penalized likelihood, sparse data.

1. INTRODUCTION

In medical studies, the effect of independent variables on dichotomous outcomes is often quantified by odds ratios that are estimated using logistic regression. Examples of such dichotomous outcomes that will be dealt with later in this paper include the development of chronic lung disease in pre-term infants (yes/no), the success in treating incontinent patients (yes/no), or the erythrocyte sedimentation rate ($> 20/\leq 20$). It is typical for medical studies that data sets are small or sparse, such that results from maximum likelihood logistic regression are not trustworthy, because odds ratio estimates are known to be biased away from one [1–5] and asymptotic confidence intervals are either not informative or violate the nominal coverage rates [6, 7]. In an extreme case, small-sample bias may cause parameter estimates to be infinite. This phenomenon has been denoted by ‘separation’, because a single independent variable or a linear combination of variables perfectly predicts the dichotomous outcome [8, 9]. Although the log likelihood converges to some finite value, it cannot be maximized by a finite parameter value.

Thus, separation leads to infinite odds ratio estimates, which rarely can be assumed to be true in practice. Finding a variable perfectly predicting the outcome is in principle very desirable. In small data sets, however, we must assume that the phenomenon of separation is not due to a truly infinite odds ratio, but rather caused by random variation.

In practice, the problem of reporting infinite odds ratio estimates is often bypassed by using a different type of model (e. g., a linear instead of a logistic model), or by replacing the covariate that is causing separation by a surrogate, by transforming that variable, or even by omitting it from the final model. None of these alternatives, however, directly estimates the effect of interest or properly adjusts the effect of other covariates by that effect. In the sequel, I will not further consider these alternatives but will rather assume the situation in which a variable’s effect on the outcome must be reported in terms of a communicable, i. e. finite, odds ratio estimate, confidence interval and P -value.

To meet the special problems that logistic regression analysis of small data sets presents, special methods have been developed, including estimation and exact inference based on conditional

likelihood [10–12] and penalized maximum likelihood estimation and inference [13–15]. As these approaches supply finite odds ratio estimates and meaningful confidence limits they are useful when analyzing separated data. These methods are also important with ‘nearly separated’ data, which can loosely be defined as data in which the existence of finite parameter estimates depends on the presence of one or two particular observations. They generally provide less biased estimates and more accurate inference.

In the following section, I briefly revisit maximum likelihood logistic regression, exact conditional logistic regression, and penalized maximum likelihood logistic regression. Section 3 compares the behavior of these methods in three typical data sets with separated or nearly separated data. Based on one of these examples, it will also be shown how a recently proposed permutation test for logistic regression models can be combined with penalized maximum likelihood estimation to confirm the adequacy of penalized likelihood ratio tests. A general discussion is given in the last section, along with information about software implementing these methods.

2. METHODS

2.1. Maximum likelihood

Consider the logistic regression model

$$\text{Prob}(y_i = 1 \mid x_i, \beta) = \pi_i = \{1 + \exp(-\sum_{r=1}^k x_{ir}\beta_r)\}^{-1}$$

with $i = 1, \dots, n$, $y_i \in \{0, 1\}$ denoting the binary outcome variable, $x_{i1} = 1$ denoting the constant, and x_{ir} ($i = 1, \dots, n$; $r = 2, \dots, k$) referring to n observations on $k - 1$ independent covariates. Maximum likelihood estimates $\hat{\beta}_r$ of regression parameters β_r ($r = 1, \dots, k$), which can be interpreted as log odds ratio estimates, are obtained as solutions to the score equations $\partial \log L / \partial \beta_r \equiv U(\beta_r) = 0$ where $\log L$ is the log-likelihood function

$$\log L = \sum_{i=1}^n y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)$$

Confidence intervals for parameters are either obtained by normal approximation (Wald method) or by profile likelihood. If $\hat{\sigma}_r$ denotes the estimated standard error of $\hat{\beta}_r$, given by the square-

root of the r th diagonal element of the inverse Fisher information matrix, then the Wald-type $(1 - \alpha)100\%$ confidence interval for β_r is defined by $[\hat{\beta}_r + z_{\alpha/2}\hat{\sigma}_r, \hat{\beta}_r + z_{1-\alpha/2}\hat{\sigma}_r]$ with z_α denoting the α -quantile of the standard normal distribution. The profile likelihood function of β_r is obtained by maximizing the log likelihood over the parameter vector β_R ($R = \{1, \dots, k\} \setminus \{r\}$, with $A \setminus B$ denoting all elements of A except those appearing in B); and inserting fixed values for β_r . A profile likelihood (PL) $(1 - \alpha)100\%$ confidence interval for β_r is the continuous set of values β_r for which twice the difference of the maximized log likelihood and the profile likelihood at β_r does not exceed the $(1 - \alpha)100$ th percentile of the χ_1^2 -distribution. While the Wald method assumes normal sample distribution of parameter estimates, the profile likelihood method allows for asymmetric distributions.

2.2. Penalized maximum likelihood

Firth [13] proposed a general method to remove first-order bias from maximum likelihood estimates. For exponential family models, his approach involves maximizing a likelihood which is penalized by Jeffreys' invariant prior. Applying his idea to logistic regression, the penalized log likelihood becomes

$$\log L^* = \log L + 1/2 \log |I(\beta)|$$

with $I(\beta)$ denoting the Fisher information matrix evaluated at β . Penalized maximum likelihood estimates for β_r ($r = 1, \dots, k$) are obtained by replacing the score equations $U(\beta_r) = \sum_{i=1}^n (y_i - \pi_i) x_{ir} = 0$ by the modified score equations

$$\partial \log L^* / \partial \beta_r \equiv U(\beta_r)^* = \sum_{i=1}^n \{y_i - \pi_i + h_i (1/2 - \pi_i)\} x_{ir} = 0 \quad (r = 1, \dots, k)$$

where the h_i 's are the i -th diagonal elements of the 'hat' matrix $H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2}$, with $W = \text{diag}\{\pi_i(1 - \pi_i)\}$. The penalization term removes the $O(n^{-1})$ -bias from parameter estimates which is considered negligible in large samples, but can be severe with small or sparse data sets. Again, confidence intervals can be obtained by either way defined above. With separated data, however, it has been shown that the penalized likelihood can be very asymmetric, and then profile penalized likelihood (PPL) confidence intervals are preferable [14].

Hypothesis tests of the parameters of the model can be obtained by comparing the unrestricted maximized penalized log likelihood with the penalized log likelihood that results by maximizing subject to the restrictions of the null hypothesis (usually setting one parameter value to zero). The penalized likelihood ratio (PLR) statistic, which is defined as twice the difference between these two penalized log likelihoods, asymptotically follows a χ^2 distribution with degrees of freedom equal to the number of parameters to test.

The penalized maximum likelihood method for binary logistic regression has been proposed as an ideal solution to the problem of separation [14]. A similar conclusion was drawn in an investigation applying the method to multinomial logistic regression [15].

2.3. Exact conditional logistic regression

In exact conditional logistic regression, the estimate of a parameter β_r as well as corresponding inference are based on the exact null distribution of the sufficient statistic $T_r = \sum_{i=1}^n y_i x_{ir}$ of β_r , conditional on the observed vector of sufficient statistics T_R ($R = \{1, \dots, k\} \setminus \{r\}$) corresponding to all regression parameters but β_r . Efficient algorithms to evaluate these conditional distributions are implemented in the software packages LogXact [16] and SAS/PROC LOGISTIC [17]. Let t_r and t_R denote the observed values of T_r and T_R , respectively. Maximum likelihood parameter estimates of β_r are found by maximizing the conditional likelihood

$$\Pr(T_r = t_r | \beta_r, T_R = t_R) \equiv \Pr_r(T_r = t_r | \beta_r) = L_r(\beta_r | T = t) = \frac{\exp(\beta_r T_r)}{\sum_{y^* \in \Omega_r} \exp(\beta_r \sum_i (y_i^* x_{ir}))}$$

with Ω_r denoting the set of permutations y^* of the outcome vector y such that for each $y^* \in \Omega_r$, $\sum_i (y_i^* x_{ir'}) = T_{r'}$ for all $r' \in R$. In case of separation t_r is at the boundary of its support. In such cases, the maximum likelihood estimate is not finite, and can be replaced by a median unbiased estimate, which is defined as the value of β_r that satisfies $L_r(\beta_r | T = t) = 1/2$ [11]. Both programs for exact logistic regression mentioned above automatically report the median unbiased estimate in case of an infinite maximum likelihood estimate.

Several methods exist to compute P -values and estimate confidence sets for β_r based on the exact conditional distribution of T_r . The ‘conditional scores’ method first assigns a score to each

distributional point of T_r measuring its distance from the mean of the distribution. The P -value is then defined as the probability that under the null hypothesis a score greater than or equal to the score of the observed value t_r is obtained. The P -value computed by the ‘probability’ method is defined as the sum of probabilities of distributional points of T_r that are less than or equal to the probability of observing t_r under the null hypothesis. Finally, the ‘twice smaller tail’ method defines the P -value as $2 \min\{\Pr_r(T_r \leq t_r | \beta_r = 0), \Pr_r(T_r \geq t_r | \beta_r = 0)\}$. Although all of these tests could be inverted to yield confidence intervals, in SAS and LogXact confidence limits of nominal level $1 - \alpha$ are estimated by computing the values β_{r-} and β_{r+} that satisfy $\Pr_r(T_r \geq t_r | \beta_{r-}) = \alpha/2$ and $\Pr_r(T_r \leq t_r | \beta_{r+}) = \alpha/2$, respectively. This corresponds to the inversion of the ‘twice smaller tail’ test. It is not clear why the conditional score method has never been considered for computing confidence intervals, which could improve on the efficiency of interval estimation. Analysis of several examples revealed that the P -value obtained by the ‘conditional score’ method is generally smaller than that by the ‘twice smaller tail’ method. A so-called mid- P version of the confidence interval is defined by the values β_{r-} and β_{r+} satisfying $\Pr_r(T_r > t_r | \beta_{r-}) + \Pr_r(T_r = t_r | \beta_{r-})/2 = \alpha/2$ and $\Pr_r(T_r < t_r | \beta_{r+}) + \Pr_r(T_r = t_r | \beta_{r+})/2 = \alpha/2$, respectively.

An excellent review of exact conditional logistic regression can be found in the paper by Mehta and Patel [12]. Although exact logistic regression was proposed for the first time in 1970 by Cox [18], application was feasible only after the availability of reasonable computing power and the development of efficient algorithms to find the sets Ω_r [10]. In recent years, research focused on Monte Carlo approximations to the exact distributions of sufficient statistics, which again broadened the scope of problems that can be solved by exact conditional logistic regression [19].

3. EXAMPLES

Three examples are presented to illustrate the behavior of the methods introduced in the previous section. The first one is a 2×2 table with an extremely unbalanced binary outcome, an example of perhaps the most common situation of separation in practice. In the second data set, both maximum likelihood and exact conditional approaches fail. In the third example the data are

only nearly-separated, but still maximum likelihood fails and exact conditional analysis is at least questionable. Results are presented as odds ratio estimates. The odds ratios for the quantitative covariates of examples 3.2 and 3.3 refer to a change of one interquartile range. Maximum likelihood analysis and exact conditional analysis were done using SAS/PROC LOGISTIC [17]. The SAS macro FL [20] was used for penalized maximum likelihood analysis.

3.1. *Preterm infants study*

The simplest case of separation occurs in the logistic regression analysis of a two-by-two table with a zero cell count. Such a table arose in a study on pre-term infants [21] when the effect of contamination of amniotic fluid by ureaplasma urealyticum on development of chronic lung disease (CLD) in such infants was evaluated. CLD was defined as need for supplemental oxygen at 36 weeks post-conception. Among 61 pre-term infants, only four developed CLD, all of them had contaminated amniotic fluid. In comparison, contaminated amniotic fluid was found in the mothers of only 17 of 57 healthy infants without CLD. The odds ratio estimates (95% confidence limits) obtained by maximum likelihood logistic regression, penalized maximum likelihood logistic regression and exact conditional logistic regression are ∞ , 20.8 and 11.5, respectively. The odds ratio estimate by exact conditional logistic regression is a median unbiased estimate. Ninety-five per cent confidence intervals for the odds ratio by Wald, profile likelihood, profile penalized likelihood, exact and mid- P methods are $(0, \infty)$, $(3.6, \infty)$, $(2.1, 2017)$, $(1.4, \infty)$ and $(1.9, \infty)$, respectively.

3.2. *Incontinence treatment study*

In a recent paper, Potter [22] published data of an incontinence treatment study, where three physiological variables x_1 , x_2 and x_3 are assumed to influence treatment success, which was assessed eight weeks after start of treatment and occurred in 13 of 21 patients. These variables are all continuous and range from -5.6 to 2.3, from -19.8 to 27.5, and from -43 to 14, respectively. The variables approximately exhibit normal distributions and their pairwise correlation coefficients range between 0.28 and 0.48. The interesting point of that data set is that although univariate odds

ratio estimates are finite for each of the covariates, a multiple logistic regression model including all three covariates does not yield finite maximum likelihood estimates. Separation is produced by a linear combination of all three covariates. With continuous covariates, exact conditional logistic regression cannot be directly applied, because the conditional distributions of the sufficient statistics are almost always degenerate, i. e. there is only one permutation y^* that satisfies the condition that $T_R = t_R$ ($R = \{1, \dots, k\} \setminus \{r\}$) and this permutation is the observed outcome vector.

Potter shows that results of exact conditional logistic regression after categorization of the covariates highly depend on the type of categorization. This was his motivation to develop a new permutation-based test that uses the residuals of a linear regression of x_3 on x_1 and x_2 as an independent variable when evaluating the effect of x_3 on treatment success. The replacement of x_3 by the residuals e_3 leaves the corresponding parameter estimate and its standard error unchanged, but the residuals e_3 are now uncorrelated to x_1 and x_2 , which cannot be generally assumed with x_3 . The residuals e_3 are then permuted and for each permutation, a likelihood ratio P -value corresponding to the likelihood comparison of the model including e_3 , x_1 and x_2 and the model including only x_1 and x_2 is obtained. The permutation of regressor residuals (PRR) P -value for the effect of x_3 is then defined as the proportion of permuted P -values that are less than or equal to the observed asymptotic likelihood ratio P -value. The idea behind Potter's approach is that the permutation of the regressor residuals yields a null distribution of P -values corresponding to x_3 without assuming that all other parameters are zero, as would be the case if y was permuted. In all permuted data sets the sufficient statistics corresponding to x_1 and x_2 assume the same values as in the original data. Potter limited his analysis to providing a P -value, but his test could be inverted to deliver confidence intervals as well. This would need some computational effort though. Alternatively, the data can be analysed by penalized maximum likelihood logistic regression. Results by both approaches are given in Table 1. Ten thousand permutations were used to compute the PRR P -values. The columns headed 'permutation' correspond to PRR P -values from penalized likelihood ratio tests ('PLR P -value') and likelihood ratio tests ('LR P -value'), respectively.

Comparing asymptotic and permutational P -values it can be seen that the discrepancy is much higher with the likelihood ratio test than with the penalized likelihood ratio (PLR) test. From the almost perfect agreement of the asymptotic and permuted PLR P -values it can be concluded that the asymptotic penalized likelihood ratio test holds the nominal size even in such small samples. In the parameter estimate corresponding to x_1 there is a large discrepancy between the two permutation-based P -values. For this variable, the distribution of the standard likelihood ratio test P -values departs heavily from the uniform, and this causes some doubt in the validity of the permutation-based likelihood ratio P -value.

3.3. *Erythrocyte sedimentation rate study*

Collett (Reference [23], p. 9; see also Reference [24]) reports a study in which the erythrocyte sedimentation rate (ESR) was regressed on fibrinogen and γ -globulin. The ESR is the rate at which red blood cells settle out of suspension in blood under standard conditions and is used as indicator for infections and certain diseases. The ESR for a healthy individual should be below 20 mm/hr; the absolute value of the ESR is relatively unimportant. Collett suggested to remove two outliers (Reference [23], pp. 8 and 168) before fitting a logistic regression model of categorized ESR (< 20 vs. ≥ 20) on fibrinogen and γ -globulin. There are only four observations out of 30 that exhibit an ESR ≥ 20 . Fibrinogen and γ -globulin are measured on a quantitative scale, ranging from 2.15 to 5.06 and from 28 to 46, respectively. As the values of γ -globulin are integers, the exact conditional distribution of the sufficient statistic of fibrinogen is not degenerate, and estimation and inference are possible. However, the observed value of the sufficient statistic for fibrinogen is at the edge of its conditional distribution, and the conditional maximum likelihood estimate is infinite. The odds ratio estimates corresponding to an increase in fibrinogen of one interquartile range (0.93) while adjusting for γ -globulin are 11.2×10^9 by maximum likelihood, 61.0 by penalized maximum likelihood, and 10.7 by conditional median unbiased estimation. Associated 95% confidence intervals are $[4.1 \times 10^{-16}, 3 \times 10^{35}]$ by the Wald method, $[19.4, 1.7 \times 10^{67}]$ by profile likelihood, $[2.97, 1.9 \times 10^{10}]$ by profile penalized likelihood, $[2.14, \infty]$ by exact conditional inference and $[2.68, \infty]$ by the mid- P method.

The confidence intervals differ remarkably in their location and length. The sparse data situation leads to an implausible lower Wald confidence limit which is due to the failure of the normal approximation that is involved in its computation. The profile likelihood interval on the other hand postulates a lower limit of 19.4. The lower limits by the exact, mid- P and profile penalized likelihood methods do not differ that much. The overlap of fibrinogen distributions between the groups defined by ESR categories suggests that the odds ratio must be finite. Thus, an infinite upper limit as given by the exact and mid- P methods is not reasonable. This is somewhat relativized by the fact that none of the other methods finds an upper limit of clinical relevance.

3.4. Evaluation by simulation for examples 3.1–3.3

Exact analysis guarantees that the actual coverage rate of a confidence interval is at least equal to the nominal confidence level. Discreteness in the data, which is one requirement for the exact conditional analysis to be applicable, can substantially increase actual coverage rates, and it is desirable to quantify the potential coverage rate excess and corresponding loss in power. On the other hand, the actual coverage rates of confidence intervals estimated by asymptotic methods converge to the nominal rates only as n becomes large.

In order to obtain estimates of each method’s actual type-I error rate (size), actual coverage rate and power corresponding to the typical covariate structures observed in the examples above, a Monte Carlo study was performed. In this simulation analysis the observed covariate data of each example was used as design matrix and one-thousand new outcome vectors were sampled assuming a ‘null’ and two different ‘alternative’ models. The ‘null’ model was obtained by maximum likelihood logistic regression analysis of the original data restricting the parameter of interest to zero. ‘Alternative’ model (a) was given by the parameter estimates from exact conditional logistic regression, and model (b) by those estimated by penalized maximum likelihood. The variables of main interest were amniotic cavity fluid culture in the preterm infants study (Section 3.1), x_3 in the incontinence treatment study (Section 3.2), and fibrinogen in the erythrocyte sedimentation rate (ESR) study of Section 3.3. The linear predictors used to simulate data are given in Table 2. A full linear predictor based on the parameter estimates from the exact conditional analysis was

only available for the preterm infants study, as in the incontinence study exact conditional analysis could not be obtained at all, and in the ESR study the estimate for the intercept parameter was not defined due to a degenerate distribution of the corresponding sufficient statistic. Therefore, model (a) for the ESR study had to be redefined by sampling one-thousand outcome vectors with the same number of events ($ESR \geq 20$) as observed in the original data set.

For each data set simulated under the ‘null’ model and under the two ‘alternative’ models, two-sided ninety-five per cent confidence intervals for the parameters corresponding to the variables of main interest were computed by the Wald, profile likelihood, profile penalized likelihood, exact and mid- P conditional methods. The size of each method was estimated from the data sets simulated under the ‘null’ model by the proportion of confidence intervals excluding 0. Actual coverage rates and power of each method were estimated from the data sets simulated under each alternative model by the proportion of confidence intervals covering the true parameter value and excluding a parameter value of 0, respectively.

Table 3 shows the results on size, coverage and power. Profile likelihood is the only method in which the actual type I error rate exceeds the nominal rate in all examples. Therefore, profile likelihood is not considered when comparing power and coverage of the methods. In the preterm infants study, all compared approaches yield at least nominal coverage rates, and profile penalized likelihood and mid- P are equally powerful, followed by exact conditional analysis and the Wald method. Among the methods that do not violate nominal coverage and type-I error rates profile penalized likelihood produces the highest value for the lower limit, thus providing the most efficient interval estimate. In the incontinence treatment study, nominal coverage is yielded by Wald and profile penalized likelihood methods, and only the latter provides reasonable power. Exact conditional analysis could only be applied in 35 of 1000 simulated data sets. In all other data sets the conditional distribution was degenerate. In the ESR study, Wald, profile penalized likelihood and exact conditional likelihood had the nominal coverage rate, and among them it is again profile penalized likelihood that yields the highest power. In this simulation study size and coverage are based on two-sided evaluations, whereas assessment of power is essentially based on

the lower confidence limit. Thus, when comparing power one should keep in mind that violations of one-sided coverage and type-I-error rates cannot be completely ruled out for the Wald, profile penalized likelihood and mid- P methods.

Summarizing the simulation results, the penalized maximum likelihood method can be generally recommended because of three findings: (i) it could be applied to all data sets, (ii) it yielded nominal type I error and coverage rates in all situations studied and (iii) it was highly efficient.

4. DISCUSSION

The present paper illustrates situations of separation and near-separation in logistic regression by means of typical examples. Three approaches were compared: maximum likelihood, exact conditional logistic regression and penalized maximum likelihood. Analysis of the examples and evaluation by simulation show that the classical maximum likelihood method can badly fail even in the analysis of a simple 2×2 -table. As infinite parameter values are generally not plausible, maximum likelihood estimates are heavily biased, and confidence intervals either uninformative, if computed by the Wald method, or extremely anticonservative, if the profile likelihood method is used. One could stop here claiming that with separated data, no inference was possible. However, research revealed that the behavior of two other approaches is more promising; these methods include exact conditional logistic regression, and penalized maximum likelihood estimation.

By means of analysis of data sets some limitations of the exact conditional logistic regression method are illustrated, which are not shared by the penalized maximum likelihood method. The first one is applicability: with continuous covariates, the conditional distributions of sufficient statistics are typically degenerate, and then neither estimation nor inference is possible. Categorization can help; however, it was remarked that results may depend heavily on the type of categorization chosen [22]. Penalized maximum likelihood estimation, on the other hand, can be readily applied to data with continuous covariates. The second problem is conservatism due to discreteness. Exact confidence intervals guarantee that the actual coverage rate is at least equal to the nominal confidence level. However, the actual coverage rate may be much higher than the

nominal level, leading to confidence intervals that are too wide and thus loss of statistical power. By constructing confidence intervals based on the mid- P -method, some of this conservatism can be removed, but the actual confidence level remains uncertain, albeit its expected value now equals the nominal level. Such confidence intervals are still based on the technique of conditioning out nuisance parameters and some of the conservatism left could be due to over-conditioning [25].

Penalized maximum likelihood estimation overcomes both problems. In the example of Section 3.2 it was shown that asymptotic P -values from penalized likelihood ratio tests fairly agree with their permutation-based counterparts, which is apparently not the case with P -values resulting from standard likelihood ratio tests. Evaluation by simulation revealed that the actual coverage rates of profile penalized likelihood confidence intervals approximately equal their nominal value while being efficient to detect departures from the null hypothesis.

The application of penalized maximum likelihood logistic regression is facilitated by SAS [17], R (<http://www.r-project.org>) and SPLUS [26] programs for routine use which are available at the web site <http://www.meduniwien.ac.at/msi/biometrie/programme/fl> [20, 27]. In these programs, parameter estimation is based on the Newton-Rhaphson algorithm and profile penalized likelihood confidence interval estimation on a variant of the algorithm suggested by Venzon and Moolgavkar [28]. It was recognized that naive application of these algorithms may lead to numerical problems in some data sets. The most vulnerable part appeared to be the estimation of the profile penalized likelihood confidence interval for the intercept parameter. As this confidence interval is not needed in the majority of applications, it is not computed by default in our routine-use programs. Almost all of the other numerical problems that arose in applications could be solved by restricting the maximum step size allowed during one Newton-Rhaphson iteration. Other techniques that improve numerical stability and speed-up convergence include standardization and orthogonalization of covariates, and ridging (see the chapter about PROC LOGISTIC in the online documentation of SAS [17] at the web site <http://support.sas.com/documentation/onlinedoc/sas9doc.html> for details). The penalized maximum likelihood approach is only slightly more complex computationally than standard maximum likelihood analysis. When doing one hundred repeated analyses

of the ESR data set on a 2.5 GHz Pentium IV machine with 1 GB of RAM running Windows XP, SAS/PROC LOGISTIC took 1.8 seconds to compute maximum likelihood estimates and profile likelihood confidence intervals, but 16.9 seconds to compute conditional maximum likelihood estimates and exact or mid- P confidence intervals. In comparison, SAS/PROC IML needed 2.1 seconds to compute penalized maximum likelihood estimates and profile penalized likelihood confidence intervals. Penalized maximum likelihood logistic regression has already been applied for the analysis of various medical studies [21, 29–33]. These studies and others, which have not yet been published, and the analysis of the examples provided by the present paper confirm the penalized maximum likelihood approach to be an easy-to-use method in the analysis of logistic regression problems when conventional asymptotic methods are in doubt and exact results are unavailable or unnecessarily conservative.

ACKNOWLEDGEMENT

I thank Alan Agresti, Paul Eilers, Michael Schemper, Harry Southworth, Ian White, Alexandra Kaider and three anonymous referees for useful comments and suggestions.

REFERENCES

1. Schaefer RL. Bias correction in maximum likelihood logistic regression. *Statistics in Medicine* 1983; **2**:71–78.
2. Cordeiro GM, McCullagh P. Bias correction in generalized linear models. *Journal of the Royal Statistical Society B* 1991; **53**:629–643.
3. Leung DH-Y, Wang YG. Bias reduction using stochastic approximation. *Australian & New Zealand Journal of Statistics* 1998; **40**:43–52.
4. Cordeiro GM, Cribari-Neto F. On bias reduction in exponential and non-exponential family regression models. *Communications in Statistics – Simulation and Computation* 1998; **27**:485–500.
5. Bull SB, Greenwood, CMT, Hauck, WW. Jackknife bias reduction for polychotomous logistic regression. *Statistics in Medicine* 1997; **16**:545–560.

6. Hauck WW, Donner A. Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association* 1977; **72**:851–853.
7. Heinze G. The impact of separation in logistic regression on maximum likelihood and other estimation methods. Unpublished Doctoral Thesis, University of Vienna, Vienna, Austria, 1998.
8. Day N, Kerridge DF. A general maximum likelihood discriminant. *Biometrics* 1967; **23**:313–328.
9. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 1984; **71**:1–10.
10. Hirji KF, Mehta CR, Patel NR. Computing distributions for exact logistic regression. *Journal of the American Statistical Association* 1987; **82**:1110–1117.
11. Hirji KF, Tsiatis AA, Mehta CR. Median unbiased estimation for binary data. *The American Statistician* 1989; **43**:7–11.
12. Mehta CR, Patel NR. Exact logistic regression: theory and examples. *Statistics in Medicine* 1995; **14**:2143–2160.
13. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; **80**:27–38.
14. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Statistics in Medicine* 2002; **21**:2409–2419.
15. Bull S, Mak C, Greenwood CMT. A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis* 2002; **39**:57–74.
16. Cytel Software Corporation. *LogXact-5*. Cytel Software Corporation, Cambridge, MA, USA, 2002.
17. SAS Institute Inc. *SAS for Windows, Version 9*. SAS Institute Inc., Cary, NC, USA, 2003.
18. Cox DR. *Analysis of Binary Data*, Methuen: London, 1970.
19. Mehta C, Patel N, Senchaudhuri P. Efficient Monte Carlo methods for conditional logistic regression. *Journal of the American Statistical Association* 2000; **95**:99–108.
20. Heinze G, Ploner M. Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. *Computer Methods and Programs in Biomedicine* 2003; **71**:181–187.
21. Berger A, Witt A, Haiden N, Kretzer V, Heinze G, Kohlhauser C. Microbial invasion of the amniotic cavity at birth is associated with adverse short-term outcome of preterm infants. *Journal of Perinatal Medicine* 2003; **31**:115–121.

22. Potter DM. A permutation test for inference in logistic regression with small- and moderate-sized data sets. *Statistics in Medicine* 2005; **24**:693–708.
23. Collett D. *Modelling binary data (2nd edition)*, Chapman and Hall: London, 2003.
24. King EN, Ryan TP. A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. *The American Statistician* 2002; **56**:163–170.
25. Pierce DA, Peters D. Improving on exact tests by approximate conditioning. *Biometrika* 1999; **86**:265–277.
26. Insightful Corp. *S-PLUS for Windows, Version 6.2*, Insightful Corp., Seattle, WA, USA, 2003.
27. Heinze G, Ploner M. Technical Report 2/2004: A SAS Macro, S-Plus Library and R Package to Perform Logistic Regression Without Convergence Problems, Department of Medical Computer Sciences, Medical University of Vienna, Vienna, 2004.
28. Venzon DJ, Moolgavkar SH. A method for computing profile-likelihood based confidence intervals. *Applied Statistics* 1988; **37**:87–94.
29. Gulyaeva N, Zaslavsky A, Lechner P, Chlenov M, McConnell O, Chait A, Kipnis V and Zaslavsky B. Relative hydrophobicity and lipophilicity of drugs measured by aqueous two-phase partitioning, octanol-buffer partitioning and HPLC. A simple model for predicting blood-brain distribution. *European journal of Medicinal Chemistry* 2003; **38**:391–396.
30. Chauchemez B, Extramiana F, Cauchemez S, Cosson S, Zouzou H, Meddane M, D’Allonnes LR, Lavergne T, Leenhardt A, Coumel P and Houdart E. High-flow perfusion of sheaths for prevention of thromboembolic complications during complex catheter ablation in the left atrium. *Journal of Cardiovascular Electrophysiology* 2004; **15**:276–283.
31. Bonderman D, Jakowitsch J, Adlbrecht C, Schemper M, Kyrle PA, Schoenauer V, Exner M, Klepetko W, Kneussl MP, Maurer G, Lang I. Medical conditions increasing the risk of chronic thromboembolic pulmonary hypertension, *Thrombosis and Haemostasis* 2005; **93**:512–516.
32. Gorak E, Geller N, Srinivasan R, Espinoza-Delgado I, Donohue T, Barrett AJ, Suffredini A, Childs R. Engraftment syndrome after nonmyeloablative allogeneic hematopoietic stem cell transplantation: Incidence and effects on survival. *Biology of Blood and Marrow Transplantation* 2005; **11**:542–550.
33. Heintel D, Kienle D, Shehata M, Kroeber A, Kroemer E, Schwarzinger I, Mitteregger D, Le I, Gleiss A, Mannhalter C, Chott A, Schwarzmeier J, Fonatsch C, Gaiger A, Doehner A, Stilgenbauer S, Jaeger

U. High expression of lipoprotein lipase in poor risk b-cell chronic lymphocytic leukemia. *Leukemia* 2005; **19**:1216–1223.

Table 1: Analysis of incontinence treatment study

Variable	PML analysis			PLR P -value		LR P -value	
	IQR	OR	95% CI	asymptotic	permutation	asymptotic	permutation
x_1	2.6	0.11	$(6.7 \times 10^{-4}, 2.27)$	0.199	0.173	8.5×10^{-4}	0.092
x_2	8.6	0.057	$(7.8 \times 10^{-9}, 0.62)$	0.0029	0.0011	2.4×10^{-6}	0.0006
x_3	11.0	3.25	(1.06, 117)	0.036	0.023	1.5×10^{-4}	0.026

PML: penalized maximum likelihood

PLR: penalized likelihood ratio

LR: likelihood ratio

IQR: interquartile range

OR: odds ratio estimate referring to an increase of one IQR

95% CI: nominal 95% confidence interval

Table 2: Linear predictors used for evaluation by simulation based on several examples. The linear predictors were obtained by restricted maximum likelihood analysis for the null models, exact conditional analysis for alternative (a) models, and penalized maximum likelihood analysis for alternative (b) models.

Example	Model	Linear predictor
Preterm infants study	null	-2.6568
Preterm infants study	alternative (a)	$-4.0467 + 2.4452 x^*$
Preterm infants study	alternative (b)	$-4.39445 + 3.03633 x^*$
Incontinence treatment study	null	$-1.1709 - 0.6952 x_1 - 0.3356 x_2$
Incontinence treatment study	alternative (b)	$0.23343 - 0.85274 x_1 - 0.33284 x_2 + 0.10708 x_3$
ESR study	null	$-4.7667 + 0.0806 \gamma\text{-Globulin}$
ESR study	alternative (a)	$2.5491 \text{ Fibrinogen} + 10^{-16} \gamma\text{-Globulin}^\dagger$
ESR study	alternative (b)	$-17.42 + 4.4209 \text{ Fibrinogen} + 0.0497 \gamma\text{-Globulin}$

* $x = 1$ if amniotic cavity culture found, 0 otherwise

† restricted to four events per simulated data set

Table 3: Results from evaluation by simulation based on several examples. The variables of interest for each example are given in brackets. The linear predictors used to simulate data were obtained by restricted maximum likelihood analysis for the null models, exact conditional analysis for alternative (a) models, and penalized maximum likelihood analysis for alternative (b) models.

Model	Preterm infants (Amniotic cavity culture)					Incontinence (x_3)			ESR (Fibrinogen)				
	null	alternative (a)		alternative (b)		null	alternative (b)		null	alternative (a)		alternative (b)	
Method	size%	cov%	pow%	cov%	pow%	size%	cov%	pow%	size%	cov%	pow%	cov%	pow%
Wald	0.5	95.9	12.0	96.0	15.8	0.0	98.8	0.2	0.8	97.0	22.7	97.6	21.3
PL	6.3	94.8	63.1	96.7	78.6	17.8	67.6	66.4	10.9	82.8	90.8	78.2	99.4
PPL	2.6	95.6	53.6	95.6	72.6	2.9	96.9	41.1	4.5	95.2	87.3	95.4	99.2
Exact	0.6	97.8	34.1	98.7	53.4	*	†	†	4.1	96.2	83.4	98.0	90.0
mid- P	2.6	96.8	53.6	97.4	72.6	*	†	†	4.8	96.2	84.5	97.9	95.8

size%: type-I-error probability; percentage of nominal 95% confidence intervals excluding the true value of 0

cov%: percentage of nominal 95% confidence intervals including the true parameter value

pow%: percentage of nominal 95% confidence intervals excluding 0

PL: profile likelihood

PPL: penalized profile likelihood

* values unavailable because only 14 of 1000 simulated data sets could be analysed

† values unavailable because only 35 of 1000 simulated data sets could be analysed